SIAM Data Mining "Brings It" to Annual Meeting

Jeremy Kepner[1] (Data Mining SIAG Chair), Sanjukta Bhowmick[2], Aydın Buluç[3], Rajmonda Caceres[1], Jordan Crouser[4], Vijay Gadepally[1], Ben Miller[1], Jennifer Webster[5]

[1]MIT Lincoln Laboratory, [2]University of Nebraska Omaha, [3]Lawrence Berkeley National Laboratory, [4]Smith College, [5]Pacific Northwest National Laboratory

The Data Mining Activity Group is one of SIAM's most vibrant and dynamic activity groups. To better share our enthusiasm for data mining with the broader SIAM community, our activity group organized six minisymposia at the 2016 Annual Meeting. These minisymposia included 48 talks organized by 11 SIAM members on

- GraphBLAS (Aydın Buluç)
- Algorithms and statistical methods for noisy network analysis (Sanjukta Bhowmick & Ben Miller)
- Inferring networks from non-network data (Rajmonda Caceres, Ivan Brugere & Tanya Y. Berger-Wolf)
- Visual analytics (Jordan Crouser)
- Mining in graph data (Jennifer Webster, Mahantesh Halappanavar & Emilie Hogan)
- Scientific computing and big data (Vijay Gadepally)

These minisymposia were well received by the broader SIAM community, and below are some of the key highlights.

GraphBLAS

The theory of using matrices and vectors for graph computations has a long history, with a snapshot of the state-of-the-art being captured in the SIAM book *Graph Algorithms in the Language of Linear Algebra* by Kepner and Gilbert [1]. High-performance graph algorithms are often implemented with sparse matrices and linear algebra in many graph-processing systems. Example systems include the Combinatorial BLAS [2], D4M [3], GraphMat [4], and GPI [5]. The GraphBLAS.org [6] is a community initiative to standardize these different efforts to build a common foundation for graph algorithm developers. This minisymposium had 8 talks: Aydın Buluç from Lawrence Berkeley National Laboratory talked about the current status of the C language API and the

This material is based in part upon work supported by the NSF under grant number DMS-1312831, by DOE ASCR under contract number DE-AC02-05CH11231, and by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

ongoing efforts to develop a GraphBLAS-compliant parallel library in PGAS (partitioned global address space) languages. Jose Moreira and Manoj Kumar from IBM presented the Graph Programming Interface (GPI) as well as a proposal for a common binary format for storing graphs. Carl Yang from UC Davis talked about implementing breadth-first search utilizing the GraphBLAS primitives on clusters of GPU-equipped computers. Andrew Lumsdaine from Indiana University talked about the software and systems issues related to implementing the GraphBLAS Template Library (GBTL) [7] on different backends, such as CPUs and GPUs. Jeremy Kepner from MIT Lincoln Laboratory presented the mathematical foundations of the GraphBLAS [8], with an emphasis on incidence matrices as a preferred representation for graphs in databases. Scott McMillan from the CMU Software Engineering Institute dived deeper into the details of the GBTL library, with a focus on its frontend design. Narayanan Sundaram from Intel presented GraphMat (and its distributed cousin GraphPad), which is a highly-optimized graph library whose frontend is based on vertex programming and whose backend is based on linear algebra operations. Finally, Michael Wolf from Sandia National Laboratories presented miniTri [9], a triangle enumeration-based data analytics miniapp, with specific focus on a linear algebraic algorithm (though miniTri has alternative algorithms in it). The sessions had lively discussions during breaks and were attended by approximately 25 people.

Algorithms and Statistical Methods for Noisy Network Analysis

Dealing with errors and noise is a common problem that the network science research community is beginning to address. A two-part minisymposium demonstrated the diversity of approaches to this problem, focusing on statistical methods and algorithms for addressing issues arising from noise in networks. Several presentations discussed useful properties of networks, such as centrality metrics and connected components, and the ways noise in the observations can affect the analysis [10]. These talks included generative models for networks and statistically rigorous methods to estimate properties from sampled data [11]. Other talks focused on filtering techniques, such as using metadata to narrow a search from a cue vertex or emphasizing an interesting substructure [12]. Some speakers discussed how noise affects the analysis in specific disciplines including collaboration science [13], bioinformatics [14], and cybersecurity. The minisymposium concluded with a discussion among the speakers and the audience on the common themes that arose. Participants agreed that, as noisy network analysis continues to evolve as a subfield, addressing the lack of a common framework for modeling and quantifying noise is an exceptionally important challenge that would allow synthesis of related research in many diverse areas.

Inferring Networks from Non-Network Data

This minisymposium explored the important topic of network representation learning. In many practical settings, researchers are faced with having to make arbitrary decisions on how to construct networks from noisy, indirect, and diverse

data. Papers presented on both sessions covered important highlights from the current state-of-art for this emerging research area. Several speakers discussed the importance of connecting the objective of a learning task, whether that is link prediction, diffusion estimation, or vertex classification, to the process of constructing and evaluating network representations [15-19]. Another important theme emphasized domain-specific notions of quality, for example, in the context of constructing robust correlation networks from biological and climate data [20, 21]. Overall, the minisymposium helped consolidate important ideas, insights, and perspectives aimed at developing a rigorous and cohesive framework for learning robust network representations.

Scientific Computing and Big Data

This two-part minisymposium was a great success. We had nine speakers from diverse organizations that shared their considerable experience working with scientific big data. In the first session, we heard from Dr. Vijay Gadepally (MIT), Dr. Siddharth Samsi (MIT), Dr. Manoj Kumar (IBM Research), Dr. Michel Kinsy (Boston University), and Dr. Shashank Yellapantula (GE Global Research). Dr. Gadepally and Dr. Samsi discussed advances in data management technologies [22–25], and Dr. Kumar presented a brief overview of a graph-based API IBM is developing [26]. Dr. Kinsy discussed a novel processing architecture for low power computations [27]. Finally, Dr. Yellapantula discussed GE's big data problems and many potential areas of collaboration with the wider SIAM community [28]. During the second session, we heard from a number of people in the medical community. Dr. Ashok Krishnamurthy (RENCI, UNC Chapel Hill) presented their development of a large-scale clinical data warehouse at the University of North Carolina Health Center [29]. Dr. Steve Finkbeiner (UCSF, Gladstone Institute) presented his group’s development of new robotic sensors capable of generating terabytes of imaging data per day to better understand the affects and causes of amyotrophic lateral sclerosis (ALS) [30]. Dr. Andy Zimolzak (Harvard, Department of Veteran Affairs) discussed his group’s work in developing computational infrastructure for precision oncology [31]. Dr. Aaron Elmore (University of Chicago) concluded the second session by presenting a new tool his research team is developing to be the GitHub for data – DataHub [32]. The presentations were of great interest to the diverse audience and there were many interesting discussions during the two sessions. Overall, the speakers and participants were left with a greater understanding of some domain-specific problems and technical strategies for addressing such problems.

Mining in Graph Data

Our minisymposium on Mining in Graph Data started off with organizer Jennifer Webster presenting an overview of the topic. Dr. Webster covered some common issues, including the use of found data that can be messy and the bias introduced by translating real-world problems into the language of mathematics. She highlighted these issues with examples drawn from shipping networks. Following that presentation, a second organizer, Mahantesh Halappanavar, presented algorithms for large-scale community detection. In particular he

described his parallel implementation of the Louvain modularity maximization method. The convergence results showed close agreement with the serial implementation, but the speed-up on multiple processors was significant. His group tested graphs with up to 50 million vertices and 2 billion edges. This was joint work conducted with Ananth Kalyanaraman. Our third speaker, Jevin West, spoke on mining information from citation networks. He presented "the map equation," which is based on dynamics of movement in a network and is used to discover communities based on those dynamics. A demo of his software was presented, along with a discussion of how this method can be used to discover the time evolution of communities. The final speaker of the morning session was Kamesh Madduri, who discussed a matrix factorization method for evaluating network community structure. When given a graph and a set of communities, he uses a non-negative matrix factorization to discover the relative importance of communities. One advantage of this work is that it can accommodate overlapping communities. These four talks rounded out the morning session, and we had steady attendance around 40 in the audience for all talks.

The minisymposium continued in the afternoon with Dr. David Haglin discussing (in his words “ranting about”) the many situations in which hyper-multi-graphs can be used and the current algorithmic and computational resource limitations to the analysis of such graphs. Dr. Haglin gave several examples of graphs in cyber and social networks, especially those where non-numeric edge information arises and where the graphs created become extremely large. Dr. Sanjukta Bhowmick then discussed her metrics for community permanence that aid in the mitigation of the noise present in real-world graphs. The permanence metric performed well across a variety of benchmark graphs and real-world data sets, and showed the stability of communities. We then saw Dr. Robert Bridges’ use of graph analysis techniques in the location of anomalous cyber activity as well as the more friendly changes in American football conferences. Dr. Bridges’ methods dealt with time-varying graphs, noisy data, and a host of other challenges in the generation, creation, and analysis of these graphs. The final talk of the minisymposium, given by Ariful Azad, was on comparing communities across graphs. When given two related graphs with communities identified, one might ask how the communities compare across those graphs. Azad gave examples of graphs created from biological data, such as MRI scans and also image segmentation over time. The Mixed Edge Cover (MEC) algorithm was used to match corresponding communities, and experimental results were given in these example data sets to show algorithm performance.

Visual Analytics

This minisymposium was organized by Prof. R. Jordan Crouser of Smith College. Visual analytics is “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas & Cook, 2006)[33] and is rapidly gaining ground as an important discipline complementary to applied mathematics. The two-session series featured speakers from Smith College, WPI, Bucknell, DePaul University, Washington University, MIT Lincoln Laboratory, and IBM Research, and covered

topics ranging from the design and evaluation of visual analytics systems to the role of human perception in data analysis.

References

[1] Kepner, J., & Gilbert, J. (Eds.). (2011). *Graph algorithms in the language of linear algebra*. Society for Industrial and Applied Mathematics.

[2] Buluç, A., & Gilbert, J. R. (2011). The Combinatorial BLAS: Design, implementation, and applications. *The International Journal of High Performance Computing Applications*, *25*(4), 496-509.

[3] Kepner, J., Arcand, W., Bergeron, W., Bliss, N., Bond, R., Byun, C., ... & McCabe, A. (2012, March). Dynamic distributed dimensional data model (D4M) database and computation system. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (pp. 5349-5352). IEEE.

[4] Sundaram, N., Satish, N., Patwary, M.M.A., Dulloor, S.R., Anderson, M.J., Vadlamudi, S.G., Das, D. and Dubey, P., (2015). GraphMat: High performance graph analytics made productive. *Proceedings of the VLDB Endowment*, *8*(11), 1214-1225.

[5] Ekanadham, K., Horn, W. P., Kumar, M., Jann, J., Moreira, J., Pattnaik, P., Serrano, M., Tanase, G. & Yu, H. (2016, May). Graph programming interface (GPI): a linear algebra programming model for large scale graph computations. In *Proceedings of the ACM International Conference on Computing Frontiers* (pp. 72-81). ACM.

[6] Mattson, T., Bader, D., Berry, J., Buluc, A., Dongarra, J., Faloutsos, C., Feo, J., Gilbert, J., Gonzalez, J., Hendrickson, B. and Kepner, J., (2013, September). Standards for graph algorithm primitives. In *High Performance Extreme Computing Conference (HPEC), 2013 IEEE* (pp. 1-2). IEEE.

[7] Zhang, P., Zalewski, M., Lumsdaine, A., Misurda, S., & McMillan, S. (2016, May). GBTL-CUDA: Graph Algorithms and Primitives for GPUs. In *Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International* (pp. 912-920). IEEE.

[8] Kepner, J., Aaltonen, P., Bader, D., Buluç, A., Franchetti, F., Gilbert, J., Hutchison, D., Kumar, M., Lumsdaine, A., Meyerhenke, H. and McMillan, S., (2016, September). Mathematical foundations of the GraphBLAS. In *High Performance Extreme Computing Conference (HPEC), 2016 IEEE* (pp. 1-9). IEEE.

[9] Wolf, M. M., Edwards, H. C., & Olivier, S. L. (2016, September). Kokkos/Qthreads task-parallel approach to linear algebra based graph analytics. In *High Performance Extreme Computing Conference (HPEC), 2016 IEEE* (pp. 1-7). IEEE.

[10] Segarra, S. and Ribeiro, A. (2016). Stability and continuity of centrality measures in weighted graphs. *IEEE Transactions on Signal Processing*, 64(3), 543–555. IEEE.

[11] Ganguly, A. and Kolaczyk, E. (2017) Estimation of vertex degrees in a sampled network. *arXiv preprint arXiv:1701.7203*.

[12] Smith, S., Caceres, R., Senne, K., McMahon, M., and Greer, T. (2017). Network Discovery Using Content and Homophily. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (to appear). IEEE.

[13] Bliss, N.T., Peirson, B. R. E., Painter, D., and Laubichler, M. D. (2014). Anomalous subgraph detection in publication networks: Leveraging truth. In *Signals, Systems and Computers, 2014 48th Asilomar conference on* (pp. 2005–2009). IEEE.

[14] Dempsey, K., Chen, T.-Y., Srinivasan, S., Bhowmick, S., Hesham, A. (2013). A structure-preserving hybrid-chordal filter for sampling in correlation networks. In *High Performance Computing and Simulation (HPCS), 2013 International Conference on* (pp. 243–250).

[15] Gleich, D. F., Mahoney, M. W (2015). Using local spectral methods to robustify graph-based learning algorithms. *In Proc. of the 21th ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining (p 359-368).*

[16] Fish. B, Caceres, R. S. (2015). Handling oversampling in dynamic networks using link prediction. *In Joint European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD) (p 671-686).*

[17] D'Amour, A., Airoldi, E. Misspecification, Sparsity, and Superpopulation Inference for Sparse Social Networks, *2016.*

[18] De, A., Valera, I., Ganguly, N.,Bhattacharya, S.,Rodriguez, M. G. Learning and Forecasting Opinion Dynamics in Social Networks, *Advances in Neural Information Processing Systems, 2016.*

[19] Brugere, I., Latent Graph Inference and Validation, *IEEE Inter. Conf. on Data Mining Workshop (ICDMW), 2015*.

[20] Agrawal, S., Atluri, G., Liess, S., Chatterjee, S.,Kumar, V. Tripoles: A New Class of Climate Teleconnections, Technical Report, 2015.

[21] Marbach, D., Lamparter D., Quon, G., Kellis M., Kutalik, Z., Bergmann, S. Tissue-Specific Regulatory Circuits Reveal Variable Modular Perturbations Across Complex Diseases. *Nature Methods 13, 2016, (p 366–370).*

[22] Elmore, A., Duggan, J., Stonebraker, M., Balazinska, M., Cetintemel, U., Gadepally, V., ... & Madden, S. (2015). A demonstration of the bigdawg polystore system. *Proceedings of the VLDB Endowment*, *8*(12), 1908-1911.

[23] Gadepally, V., Chen, P., Duggan, J., Elmore, A., Haynes, B., Kepner, J., ... & Stonebraker, M. (2016, September). The BigDAWG polystore system and architecture. In *High Performance Extreme Computing Conference (HPEC), 2016 IEEE* (pp. 1-6). IEEE.

[24] Samsi, S., Brattain, L., Arcand, W., Bestor, D., Bergeron, B., Byun, C., ... & Michaleas, P. (2016, September). Benchmarking SciDB data import on HPC systems. In *High Performance Extreme Computing Conference (HPEC), 2016 IEEE* (pp. 1-5). IEEE.

[25] Hutchison, D., Kepner, J., Gadepally, V., & Fuchs, A. (2015, September). Graphulo implementation of server-side sparse matrix multiply in the Accumulo database. In *High Performance Extreme Computing Conference (HPEC), 2015 IEEE* (pp. 1-7). IEEE.

[26] Horn, W., Kumar, M., Jann, J., Moreira, J., Pattnaik, P., Serrano, M., ... & Yu, H. (2017). Graph Programming Interface (GPI): A Linear Algebra Programming Model for Large Scale Graph Computations. *International Journal of Parallel Programming*, 1-29.

[27] Kinsy, M. A., Pellauer, M., & Devadas, S. (2011, September). Heracles: Fully synthesizable parameterized mips-based multicore system. In *Field Programmable Logic and Applications (FPL), 2011 International Conference on* (pp. 356-362). IEEE.

[28] Yellapantula, S., Venkatesan, K., Pratt, A., Slabaugh, C., & Lucht, R. P. (2016). LES validation practices in a model aero-engine combustor at engine relevant conditions. In *52nd AIAA/SAE/ASEE Joint Propulsion Conference* (p. 4785).

[29] Evans, J. P., Wilhelmsen, K. C., Berg, J., Schmitt, C. P., Krishnamurthy, A., Fecho, K., & Ahalt, S. C. (2016). A New Framework and Prototype Solution for Clinical Decision Support and Research in Genomics and Other Data-intensive Fields of Medicine. *eGEMs*, *4*(1).

[30] Lammel, G., Armbruster, S., Schelling, C., Benzel, H., Brasas, J., Illing, M., ... & Finkbeiner, S. (2005, June). Next generation pressure sensors in surface micromachining technology. In *Solid-State Sensors, Actuators and Microsystems, 2005. Digest of Technical Papers. TRANSDUCERS'05. The 13th International Conference on* (Vol. 1, pp. 35-36). IEEE.

[31] Celi, L. A., Zimolzak, A. J., & Stone, D. J. (2014). Dynamic clinical data mining: search engine-based decision support. *JMIR medical informatics*, *2*(1), e13.

[32] Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A. J., Madden, S., & Parameswaran, A. G. (2014). Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798*.

[33] Thomas, James J. and Cook, Kristin, eds. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.